DOCUMENT RESUME

ED 401 322 TM 025 902

AUTHOR Wang, Tianyou; Zeng, Lingjia
TITLE Item Parameter Estimation for the Continuous Response Model via an EM Algorithm.
PUB DATE Apr 96
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Ability; *Educational Assessment; *Estimation (Mathematics); *Item Response Theory; Psychological Testing; Simulation; *Test Items
IDENTIFIERS Calibration; EM Algorithm; *Item Parameters; Work Keys (ACT)

ABSTRACT

F. Samejima (1973) proposed a continuous response model in which item response is on a continuous scale rather than some discrete levels. This model has potential because in many psychological and educational assessments, the responses are on a conceptual continuum rather than on some fixed levels. As a first step toward studying the applicability of the continuous response model to psychological and educational assessment, it was important to develop an item parameter estimation algorithm. Because the estimation of item parameters involves an incidental theta parameter, the EM algorithm is used. The EM algorithm is a popular technique for item response theory item parameter estimation that is used in commercial calibration software. The item parameter estimation procedure is described step-by-step. The procedure was programmed in C language, and it was evaluated using real and simulated test data. The real data were from 3 forms of the Work Keys (American College Testing Program) writing test assessment with samples of 7,097, 2,035, and 1,793 people. Simulated data were also based on Work Keys writing test forms for 200, 500, and 2,000 simulated examinees. Results of both studies show that the continuous response model seems applicable to real test data with many categories of responses and that the EM algorithm works reasonably well in recovering true item parameter values. (Contains 2 tables, 3 figures, and 12 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 322

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
■ This document has been reproduced as received from the person or organization originating it.
□ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
TIANYOU WANG
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Item Parameter Estimation for the Continuous Response Model Via an EM Algorithm

Tianyou Wang and Lingjia Zeng

Measurement Research Department

ACT

A paper presented at the 1996 Annual Meeting of the American Educational Research Association, New York, April.

Address correspondence to Tianyou Wang, ACT, P.O.Box 168, Iowa City, IA 52243.

TM025902

BEST COPY AVAILABLE

Item Parameter Estimation for the Continuous Response Model Via an EM Algorithm

Tianyou Wang and Lingjia Zeng

ACT

In recent years, more research effort has been devoted to polytomous item response models due to increased uses of performance type of assessment. So far, commercial software such as PARSCALE (Muraki & Bock, 1993), Multilog (Thissen, 1991), and FACETS (Linacre & Wright, 1993) are available for calibration for various polytomous IRT models: nominal response model (Bock, 1972), rating scale model (Andrich, 1978), graded response model (Samejima, 1969), partial credit model (Masters, 1982), and generalized partial credit model (Muraki, 1992), etc. In general, these models allow parameters of difficulty levels for each response category. Some models allow the discrimination parameters to vary from item to item whereas some Rasch type models do not. Although applications of the polytomous IRT models to performance assessment data have been generally successful in many situations, polytomous IRT models also have some potential problems. First, polytomous IRT models requires a category parameter for each response category; there must be enough observations for each category in order to calibrate the category parameter. Second, in situations where scores are obtained using human raters, often more than one rater is used, and the resulting scores are usually the average of two or three ratings and will not necessarily be integer score points. Third, in some situations where ratings can potentially be any point on a continuum, forcing rating scores to fall on integer points will necessarily introduce rounding errors. Given these problems, alternative approaches in modeling performance assessment data deserve consideration.

Samejima (1973) proposed a continuous response model in which item response is on a continuous scale rather than some discrete levels. This model has some potential because in many (perhaps most) psychological and educational assessments, conceptually the responses are on a continuum rather than on some fixed levels. Even though a finite number of anchor points on a

continuous scale may facilitate the rating process, the final ratings do not have to fall on these anchor points. Rather a rating should be allowed to fall between two integer points if a rater finds the performance is between two levels. Theoretically, this practice should increase the precision of the ratings. It is true that in some assessment settings, the amount of trainings on the rating task the raters receive make it difficult for the raters to make judgments more finely than, say, on a 6 point scale. But because the quality of examinee performances are mostly on a continuum instead of at discrete levels, it is always desirable to fine tune the rating process so that the ratings are more geared to the actual examinee performances. This point is exemplified by the Olympic gymnasitic or diving races where expert judgers usually give ratings on a scale a 6 point or 10 point scale but with one decimal place. In educational measurement, ratings on a continuous scale do not pose any new challenge to the classical test theroy-based techniques because classical test theory do not assume the discrete or continuous nature of the test data. The currently available IRT-based techniques, however, mostly assume the response data are of discrete nature. Research on applying model-based measurement technique on more continuous type of response data are almost non-existent, probably due to the unavailability of a calibration program. For this reason, it seems imperative to study the applicability of the continuous response model to psychological and educational assessment. As a first step, it is important to develop an item parameter estimation algorithm.

The Continuous Response Model

Samejima (1973) presented a general form of the continuous response model as a limiting situation of the graded response model. In her formulation, the probability for an examinee with latent ability level θ of getting score of x or higher $(X \geq x)$ on a score scale ranging from 0 to k for a particular item is an increasing function of θ. For example, this function may follow a logistic curve or a normal ogive curve. Subtracting this probability function from one gives the cumulative probability density function (cdf) of response X conditioned on θ. The probability density function (pdf) of X is thus the first derivative of the cdf. It turns out that if the probability

function of $(X \geq x)$ takes a normal ogive form, it is quite easy to derive the pdf. For convenience, the derivations in this paper are all based on the normal ogive function. Let us define the probability of an examinee getting a score of x or higher on a particular item as

$$P(X \geq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-t^2/2} dt, \tag{1}$$

where,

$$u = a\left(\theta - b - \frac{\ln \frac{x}{k-x}}{\alpha} \right),$$

where a represents a discrimination parameter, and b represents a difficulty parameter, and α represents a scaling parameter which defines some scale transformation that links the raw rating scale to the latent trait θ scale. The formulation here represents a slightly different parameterization than in Samejima's article, but they are essentially the same. As a illustration of the probability function in equation 1, Figure 1 plots the the cumulative probability density on the θ scale for three hypothetical items with different a, b, and α parameters. The rating on these items ranges from 0 to 5. The four curves in Figure 1 represent the cumulative probability density at four integer points. It is seen from Figure 1 that the a parameter affects the slope of the curves; the b parameter affects the loction of the curves; the α parameter affects the distance between those curves. It is also seen that the curves are not of equal distance, but they are symmetric around the mid-point of the rating scale (in this case it is 2.5). The three parameters in this model thus all have their distinct and practical meanings that are useful for item analysis purpose.

Next, the probability density functions are derived. For convenience, let

$$Z = \ln \frac{X}{k-X}, \tag{2}$$

and find the cdf and pdf for random variable Z rather than for X,

$$
\begin{aligned}
f(z) &= \frac{dF(z)}{dz} \\
&= \frac{d(1-P(Z>z))}{dz} \\
&= \frac{d(1-P(X>x))}{dz} \\
&= \frac{a}{\sqrt{2\pi}\alpha} e^{-\left(a\left(\theta-b-\frac{z}{a}\right)\right)^2/2}
\end{aligned}
\tag{3}
$$

Because a normal ogive function is used for $P(Z>z)$, the pdf of Z turns out to be a normal density function with a mean of $\alpha(\theta-b)$ and a variance of α^2/a^2. Figure 2 plot the normal density function for the same three hypothetical items as plotted in Figure 1. The effect of different item parameters on the shape of the curves are basically the same as illustrated in Figure 1. The fact that the density function is normally distributed simplifies subsequent derivations significantly.

As derived by Samejima, the item information function

$$
\begin{aligned}
I(\theta) &= -E\left(\frac{\partial^2 \ln f(z)}{\partial\theta^2}\right), \\
&= a^2
\end{aligned}
\tag{4}
$$

which remains a constant across the θ scale. Samejima proved that with a logistic function for $P(Z>z)$, the item information function is also a constant across the θ scale. The implication of the constant information property is discussed in Samejima (1973). Equation 4 indicates that the information of the item depends only on its a parameter. Because the informatin is related to the conditional standard error of measurement, equation 4 indicate that the difficulty level of the item does not affect the measurement error for an examinee of any ability level. This is quite different from what people usually encouter in IRT analysis. This constant information property is due to the mathematical form of the density function which determine the nature of the θ scale.

Samejima also illustrated with some examples that for the same a parameter value, the information value is higher for the continuous response model than for the graded response model,

suggesting that the continuous response situation is more informative than the graded response situation. This result suggest that with some discrimination power, the fine tuning of the rating data can the precision in estimating θ .

Item Parameter Estimation Using the EM Algorithm

Because the estimation of item parameters involve an incidental parameter θ , the EM algorithm (Dempster, Laird, Rubin, 1977; Tanner, 1993) is used. The EM algorithm is a popular technique for IRT item parameter estimation and is used in commercial calibration software such as BILOG and PARSCALE. When the EM algorithm is used in IRT item parameter estimation, the θ 's are treated as missing data. The most general expression of the EM algorithm is that in the E-step,

$$\begin{aligned} Q(\varepsilon, \varepsilon^n) &= E_{\theta|\varepsilon^n, Z}(\ln p(\varepsilon|\theta, Z)) \\ &= \int_\theta \ln p(\varepsilon|\theta, Z) p(\theta|\varepsilon^n, Z) d\theta \end{aligned} \tag{5}$$

where ε represents item parameters, Z represents response data, and θ represents latent trait data in this situation. $p(\varepsilon|\theta, Z)$ is the posterior distribution of the item parameters given both the observed data and latent data. The expectation is over the posterior distribution of the latent data given the response data and the item parameter estimates from the previous EM cycle.

In the M-step, $Q(\varepsilon, \varepsilon^n)$ is maximized with respect to ε to obtain ε^{n+1}. ε^{n+1} is used in the next EM cycle to obtain the updated posterior distribution of the latent data. This iterative process keeps going on until a certain convergence criterion is reached.

Given a sample of N persons' responses to m items, $Q(\varepsilon, \varepsilon^n)$ can be more specifically expressed as

$$
\begin{aligned}
Q(\varepsilon, \varepsilon^n) &= E_{\theta|\varepsilon^n, Z}\left(\ln p(\varepsilon|\theta, Z)\right) \\
&= E_{\theta|\varepsilon^n, Z}\left(\ln \frac{p(Z|\theta, \varepsilon)p(\varepsilon)}{p(Z|\theta)}\right) \\
&= E_{\theta|\varepsilon^n, Z}\left(\ln p(Z|\theta, \varepsilon)\right) + E_{\theta|\varepsilon^n, Z}\left(\ln p(\varepsilon)\right) - E_{\theta|\varepsilon^n, Z}\left(\ln p(Z|\theta)\right) \\
&= \sum_{i=1}^{N} E_{\theta|\varepsilon^n, Z}\left(\ln p_i\left(Z_i|\theta, \varepsilon\right)\right) + \ln p(\varepsilon) - \ln p(Z|\theta) \qquad (6) \\
&= \sum_{i=1}^{N} \int_{\theta} \ln p_i\left(Z_i|\theta, \varepsilon\right) p_i\left(\theta|\varepsilon^n, Z_i\right) d\theta + \ln p(\varepsilon) - \ln p(Z|\theta)
\end{aligned}
$$

The first term of this expression is a summation of each examinee's likelihood function integrated over this examinee's posterior distribution of θ. This posterior distribution is based on this examinee's response data, the item parameter estimates from the previous EM cycle, and some prior distribution of θ which can be simply fixed as standard normal or updated using the posterior distribution from the previous EM cycle. The second term is the log of some prior distribution of the item parameters. The third term is the log of some marginal distribution of the response data which does not involve item parameters and thus can be dropped in the M-step. If we have some prior information about the distribution of the item parameters, we can include the second term in the M-step, and in doing so we are adopting a Bayesian approach of parameter estimation. But for the purposes of this study, we will drop this second term to avoid the influence of different specifications of the prior distribution of the item parameters which entails further research. To drop the second term is equivalent to specifying a uniform prior distribution for each item parameter over some valid range.

To maximize $Q(\varepsilon, \varepsilon^n)$ is to set its partial derivatives with respect to item parameters to zero and simultaneously solve the equation to obtain item parameter estimates.

$$\frac{\partial Q(\varepsilon,\varepsilon^n)}{\partial \varepsilon_j} = \sum_{i=1}^{N} \int_\theta \frac{\partial \ln p_i(Z_i|\theta,\varepsilon)}{\partial \varepsilon_j} p_i(\theta|\varepsilon^n, Z_i) d\theta$$

$$= \sum_{i=1}^{N} \int_\theta \frac{\partial \sum_{j=1}^{m} \ln p_{ij}(z_{ij}|\theta,\varepsilon_j)}{\partial \varepsilon_j} p_i(\theta|\varepsilon^n, Z_i) d\theta \qquad (7)$$

$$= \sum_{i=1}^{N} \int_\theta \frac{\partial \ln p_{ij}(z_{ij}|\theta,\varepsilon_j)}{\partial \varepsilon_j} p_i(\theta|\varepsilon^n, Z_i) d\theta = 0$$

This means that in the M-step, the equations can be solved in an item-by-item fashion.

Substituting $p_{ij}(z_{ij}|\theta,\varepsilon_j)$ with the normal density function with a mean of $\alpha_j(\theta_i - b_j)$ and a variance of α_j^2/a_j^2, we can express the derivatives with respect to each item parameter as:

$$\frac{\partial \ln p_{ij}(z_{ij}|\theta,\varepsilon_j)}{\partial a_j} = \frac{1}{a_j} - a_j\left(\theta - b_j - \frac{z_{ij}}{\alpha_j}\right)^2$$

$$\frac{\partial \ln p_{ij}(z_{ij}|\theta,\varepsilon_j)}{\partial b_j} = a_j^2\left(\theta - b_j - \frac{z_{ij}}{\alpha_j}\right) \qquad (8)$$

$$\frac{\partial \ln p_{ij}(z_{ij}|\theta,\varepsilon_j)}{\partial \alpha_j} = -a_j^2\left(\theta - b_j - \frac{z_{ij}}{\alpha_j}\right)\left(\frac{z_{ij}}{\alpha_j^2}\right) - \frac{1}{\alpha_j}$$

The integration over the posterior θ distribution can be computed numerically using Gaussian quadrature points (Stroud & Sechrest, 1966) or simply equal spaced quadrature points. With q qradrature points, the M-step equations can be expressed as:

$$\sum_{i=1}^{N} \int_\theta \frac{\partial \ln p_{ij}(z_{ij}|\theta,\varepsilon_j)}{\partial \varepsilon_j} p_i(\theta|\varepsilon^n, Z_i) d\theta$$

$$\approx \sum_{i=1}^{N} \sum_{k=1}^{q} \frac{\partial \ln p_{ij}(z_{ij}|QX_k,\varepsilon_j)}{\partial \varepsilon_j} W_{ik}(QX_k|\varepsilon^n, Z_i), \qquad (9)$$

$$= 0$$

where QX_k's are the quadrature points and $W_{ik}\left(QX_k \mid \varepsilon^n, Z_i\right)$'s are the corresponding quadrature weights.

This set of three equations can be solved simultaneously using the Newton-Raphson method iteratively to obtain the item parameter estimates for the current EM cycle. The EM iterative process terminates when item parameter estimates change very little between two cycles.

Evaluation of the Parameter Estimation Procedure

The item parameter estimation procedure described above was programmed in C language. The procedure was evaluated using both real test data and simulated test data.

With Real Test Data:: There are two main purposes of using real data to evaluate this procedure. The first purpose is to examine the estimated item parameter values and the density function of some real test items. The second purpose is to examine if the continuous response model fits some real test data. A data set which contain examinee responses to three forms of the Work Keys (ACT, 1995) writing test was used in this analysis. The sample sizes for these three forms 10cc, 11cc, and 12cc are 7097, 2035 and 1793 people, respectively. Each of the test forms contain six items each with six score categories (from 0 to 5). Each of the student responses was rated by two raters. The average of the two ratings results in 11 score categories for each item. The response data were fit to the continuous response model and item parameters were calibrated. Table 1 contains the item parameter values of the items in the three forms. The *b* parameters of these items suggest that these items are quite easy for a group of having zero mean θ distribution. This is due to the fact that the responses to items have means that are above the mid-point of the score range, which makes the means of *z* scores positive. Also, the *a* and α parameters vary considerably from item to item.

Based on the item parameters and estimated examinee θ distribution generated by the estimation program, The expected *z* score distribution can be computed for each item. The distribution for the *x* score were computed using the inverse of equation 2. The *x* scores were rounded to be integers and the discrete distributions of integer *x* score were computed using

numerical integration method. The expected raw total test score distributions were then computed using Hanson's (1994) extension of the Lord-Wingersky (1984) algorithm. The expected score distributions were plotted and compared to the observed test score distributions to examine model-data fit and the accuracy of the item parameter estimation. The extent to which the expected score distributions conform to the observed score distributions indicates how the model fits the data and how accurate the item parameters are estimated. Figure 3 contains these plots of observed and expected score distributions for the three test forms. These plots show that the expected score distributions fit the shapes of the observed distribution nicely. However, it seems that the expected distributions are less skewed than the observed score distribution. The difference in the skewness is probably due to the fact that the estimation procedure used a standard normal distribution as the prior distribution for θ in the EM algorithm. This prior distribution causes the estimated θ distribution to be less skewed than the actual θ distribution.

With Simulated Data: The advantage of using simulated data is that the true item parameters are known so that item parameter estimates can be compared to their true values. Item parameter recovery capability of the estimation procedure will be evaluated using simulated data for different sample size and different number of items. The true item parameter values in this analysis were taken from the calibrated item parameter values from the three Work Keys writing test forms used above. The 6 items in form 10cc form one test while the 12 items from form 11cc and 12cc were combined together to create another test form to get a longer test length. Item responses were simulated for these two new forms based on two different θ distributions and three different sample sizes. The two θ distributions are the standard normal distribution and an estimated θ distribution from form 10cc calibration. This estimated θ distribution is moderately negatively skewed. The three sample sizes are 200, 500, and 2000 simulees, respectively. The simulated response data were input into the calibration program and the estimated item parameters were obtained. Three indices were computed to compare the estimated and the true item parameter values: the correlation, the average (across items) absolute difference, and the average signed difference. Table 2 contains the results for these indices. The results show that the estimation

procedure can accurately recover item parameter values, especially when sample size is reasonably large. The accuracy of the estimation is best for the *b* parameters, followed by that for the *a* parameters, followed by the alpha parameters. The recovery capability is slightly affected by the nature of the θ distributions used to generate responses. More accurate estimations were obtained with the standard normal distribution. When a skewed empirical distribution is used, the estimates of the *a* and alpha parameters display a slight systematic shift. This result suggest that the EM algorithm can not fully recover the skewness of the θ distribution. An examination of the estimated θ distribution based on the simulated data confirmed this conjunction.

In summary, the results of this study shows that continuous response model seems applicable to real test data with many categories of responses and that the EM algorithm works reasonably well in recovering the true item parameter values when applied to simulated response data. Further research needs to be done to study the various aspects of the continuous response model when applied to different testing situations

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. 39, (Series B),* 1-38.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to the polytomous items.* Unpublished research note.

Linacre, J. M. & Wright, B. D. (1993). *FACETS.* MESA Press: Chicago.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement,* 8, 452-461.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 1973,* 38, 203-219.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs NJ: Prentice Hall.

Tanner, M. A. (1993). Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. Springer-Verlag: New York.

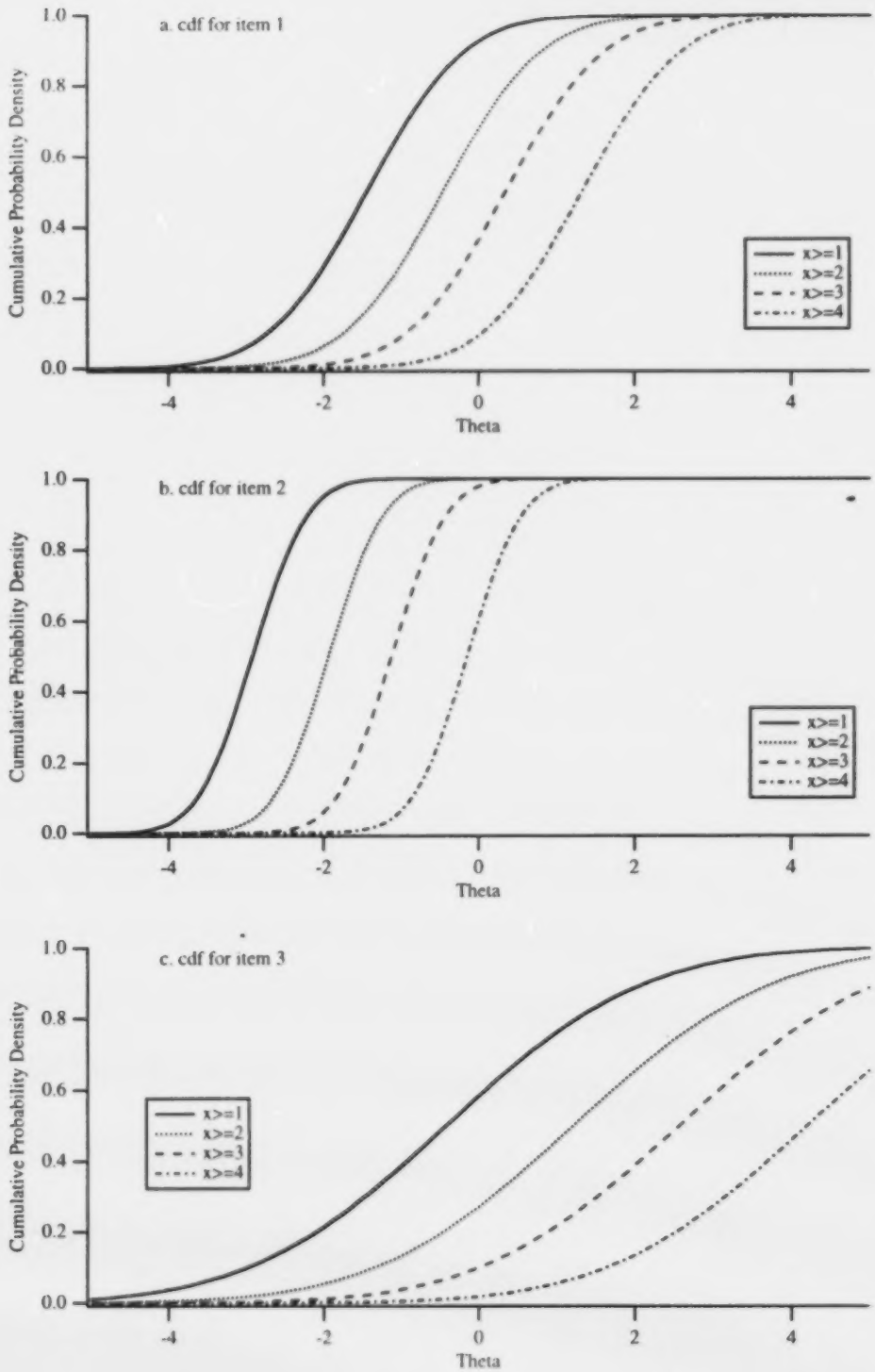

Figure 1. The cumulative probability density for three hypothetical items with parameters: item1 (a=1.0, b=0.0, alpha=1.0), item2 (a=1.8, b=-1.5, alpha=1.0) item3 (a=0.5, b=2.0, alpha=0.6).

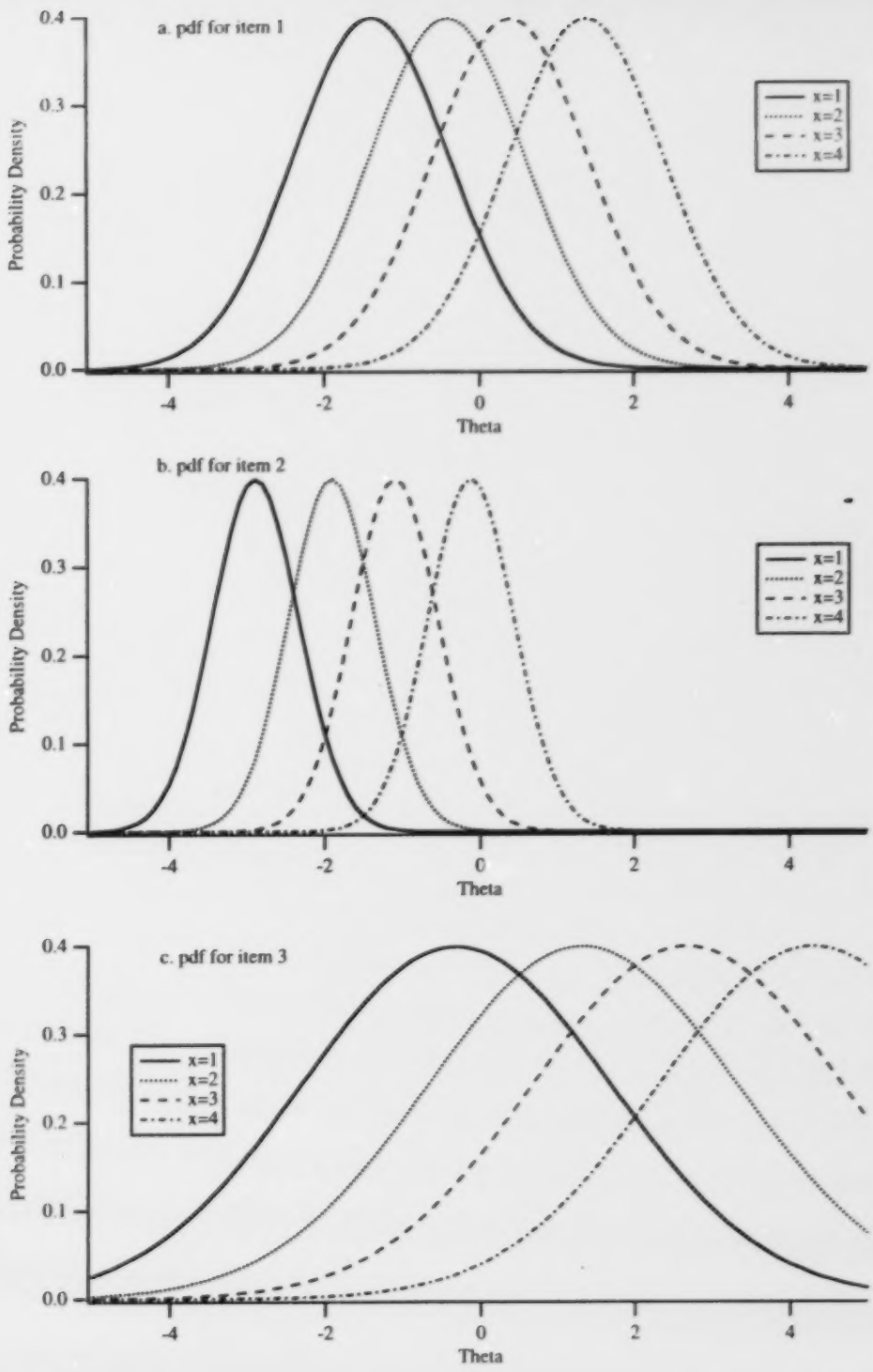

Figure 2. The probability density for three hypothetical items with parameters: item1 (a=1.0, b=0.0, alpha=1.0), item2 (a=1.8, b=-1.5, alpha=1.0) item3 (a=0.5, b=2.0, alpha=0.6).

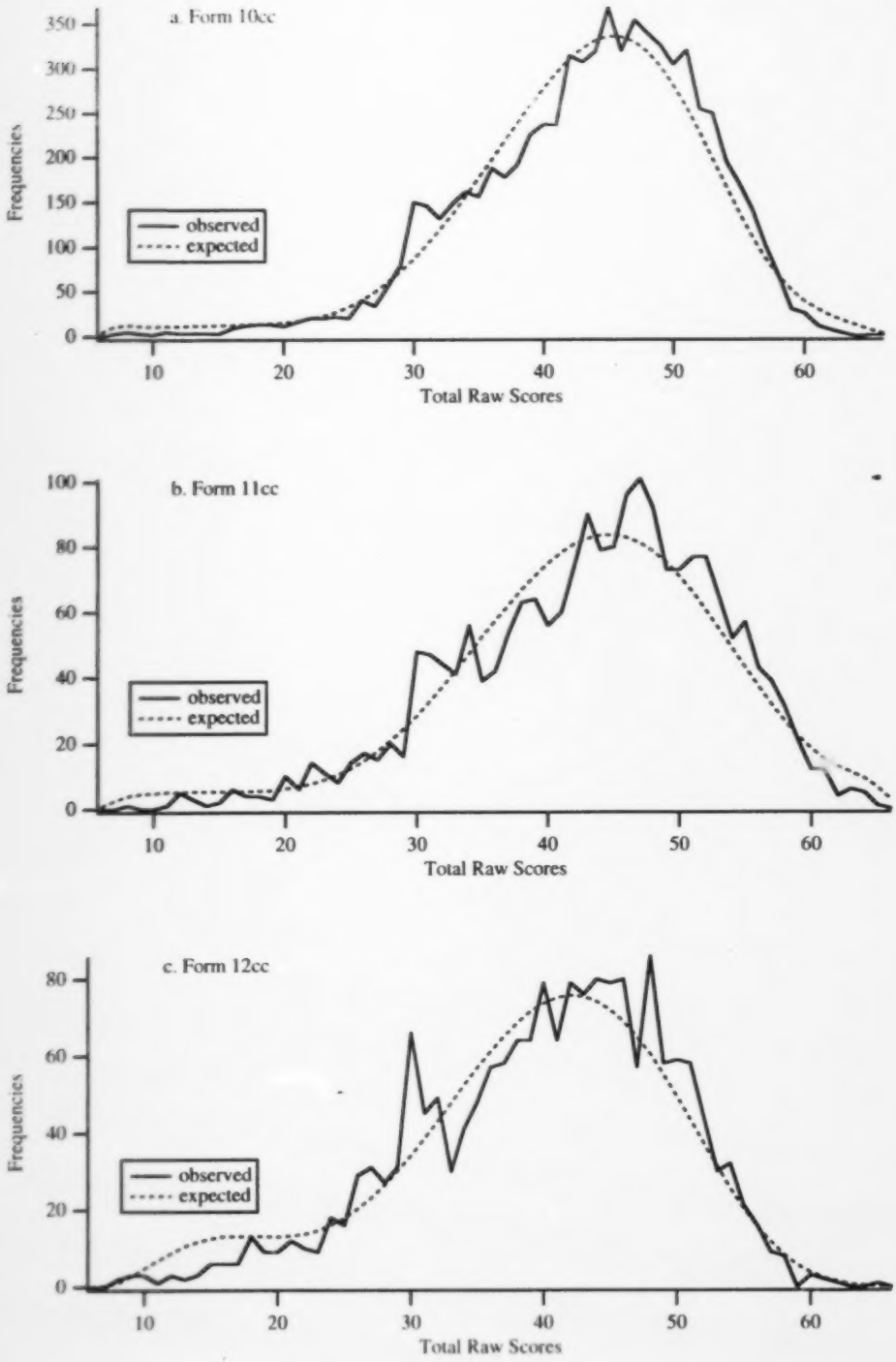

Figure 3. The fitted and observed total raw score distribions for the three Work Keys Writing forms.

Table 1. Estimated parameters for the real items from the Work Keys writing test.

	Form 10cc			Form 11cc			Form 12cc		
item	a	b	alpha	a	b	alpha	a	b	alpha
1	0.643	-1.473	0.894	0.649	-1.651	1.012	0.381	-1.119	0.470
2	0.931	-0.970	0.653	0.773	-0.934	0.725	0.361	-0.283	0.368
3	1.192	-0.720	0.733	0.990	-0.742	0.777	0.505	-1.173	0.685
4	1.432	-0.536	1.003	0.991	-0.388	0.927	1.524	-0.072	1.248
5	1.316	-0.430	1.018	1.375	-0.350	1.146	1.548	0.112	1.417
6	0.953	-0.051	1.066	0.889	-0.071	1.109	1.553	0.054	1.322

Table 2. The correlations (r), average absolute difference (aad), and average signed difference (asd) between the estimated and true item parameter values.

Ability distribution	n	r			aad			asd		
		a	b	alpha	a	b	alpha	a	b	alpha
Normal	2000	0.988	0.999	0.914	0.048	0.029	0.039	0.004	-0.014	0.017
	500	0.945	0.996	0.980	0.080	0.052	0.039	-0.016	-0.002	-0.012
	200	0.893	0.992	0.849	0.131	0.133	0.083	0.058	-0.073	0.082
Empirical	2000	0.977	0.999	0.964	0.176	0.072	0.143	0.176	0.024	0.143
	500	0.958	0.999	0.946	0.216	0.129	0.175	0.216	0.069	0.175
	200	0.984	0.996	0.975	0.131	0.095	0.125	0.131	0.084	0.125

a. For the 6-item test.

Normal	2000	0.983	0.997	0.960	0.057	0.033	0.050	-0.001	0.011	0.010
	500	0.971	0.981	0.944	0.078	0.087	0.059	0.011	0.050	0.028
	200	0.953	0.983	0.896	0.151	0.097	0.109	-0.020	0.070	0.010
Empirical	2000	0.972	0.993	0.943	0.178	0.082	0.179	0.178	0.001	0.179
	500	0.949	0.989	0.886	0.134	0.125	0.141	0.121	-0.083	0.131
	200	0.952	0.952	0.948	0.202	0.234	0.186	0.193	-0.234	0.177

b. For the 12-item test.